(NASA-CR-173804) TWO-DIMENSIONAL SPLINES (California Univ.) 59 p HC A04/MF A01 CSCL 12A

N84-29561

Unclas

G3/64 19999

NSG - 4015

UCLA ENG 84-21

TWO-DIMENSIONAL SPLINES

Florencio I. Utreras

NSG
Supported by NASA under Grant 4015, and by the Air Force Office of Scientific Research under Grant 83-0318.

UCLA • SCHOOL OF ENGINEERING AND APPLIED SCIENCE

TWO-DIMENSIONAL SPLINES*

by

Florencio I. Utreras**
Depto. de Matemáticas y Ciencias de la
Computación
Facultad de Ciencias Fisicas y Matemáticas
Universidad de Chile
Casilla 5272, Correo 3
Santaigo, CHILE

Lectures given at UCLA from June 28 to July 12, 1984

---

* Research supported in part under AFOSR Grant No. 83-0318 and NASA Grant No. 4015.

** Present address: Department of Mathematics, Texas A&M University, College Station, TX 77843

TABLE OF CONTENTS

	page
1. One-Dimensional Splines	1
1.1 The Cubic Splines	1
1.2 Smoothing Splines	9
2. Thin Plate Splines	13
2.1 Theoretical Background	13
2.2 The Computation of σ	16
2.3 A Practical Method for a Large Number of Points	22
3. A General Framework	27
3.1 Reproducing Kernels and Spline Functions	28
3.2 The Stochastic Approach	31
4. Constrained Interpolation	34
4.1 The Positive Thin Plate Spline	34
4.2 The Dual Algorithm	36
5. Thin Plate Smoothing Splines	41
5.1 Existence, Uniqueness and Characterization	41
5.2 The Computation of λ and $\sigma_{n,\lambda}$	44
6. Non-variational Techniques	48
6.1 Tensor Product Splines	48
6.2 Multidimensional B-Splines	50
6.3 Conclusion and Comments	52
References	53

1. One-dimensional Splines

The purpose of this section is to give an introduction to some outstanding properties of one-dimensional splines. We are not interested in giving a very detailed discussion of this vast subject but only in pointing out some results which serve to illustrate some properties of multivariate splines.

1.1. The Cubic Spline

Consider knots $t_1 < t_2 < t_3 < \cdots < t_n$ of a real interval $[0,T]$ and real data $y_1, y_2, \ldots, y_n$. One of the oldest problems considered by mathematicians and engineers has been that of finding a "smooth" curve joining the points $(t_i,y_i) \in \mathbb{R}^2$. Or, in mathematical terms, to find a function $g\colon [0,T] \to \mathbb{R}$ such that g takes the value y_i at t_i $i = 1,\ldots,n$.

$$g(t_i) = y_i \qquad i = 1,\ldots,n \tag{1.1.1}$$

Obviously, this problem does not have a unique solution and the choice of a solution will depend on many factors, some of which are:

- -- Smoothness properties
- -- Convergence properties
- -- Cost of computation.

Even if the first two properties have a great importance, the third one is usually the one determining the interpolant to be used. Among the solutions used in the past, the most popular one was the polynomial interpolation. As it is known, we can find a unique polynomial p_n of degree $n-1$ satisfying (1.1.1). Moreover, an explicit expression for p_n can be given in terms of Lagrange polynomials

$$p_n = \prod_{i=1}^{n} y_i \ell_i \tag{1.1.2}$$

where

$$\ell_i(x) = \frac{\prod_{i \neq j} (x - x_j)}{\omega'(x_i)} \tag{1.1.3}$$

$$\omega(x) = \prod_{i=1}^{n} (x - x_i) \quad . \tag{1.1.4}$$

Thus, the solution p_n can be easily written as a linear combination of well-known polynomials. This is a very interesting property, even if (1.1.2) is not used for practical calculations because it is cheaper to use the Newton's formula (cf. [6]).

Another interesting property of the interpolation polynomial is that it can be differentiated infinitely many times and each derivative is still a polynomial and can be easily evaluated using a Horner's scheme. This last assertion is only partly true because the evaluation of a polynomial of degree $n-1$ requires $n-1$ multiplications and for large n this could become very expensive.

However, the main trouble arising with polynomail interpolation is mainly due to the fact that they are too smooth and hence too inflexible. This leads to oscillations when the number of points is high (more than 20). Indeed, the interpolation polynomials have very bad approximation properties except for a small class of analytic functions (cf. [6]).

Schoemberg [25] found the way to avoid the problems coming from the inflexibility of polynomials and still keep their nice properties: he introduced the use of piecewise polynomials matched together by continuity conditions. He called the new class of functions: splines.

For a general description of polynomial splines the reader is referred to [1], [9], [26]. In this section we will only describe one of the most popular splines: the cubic spline.

Let S_3 be the set of all functions $[t_1,t_n]$ satisfying the following properties:

CS1: σ is a cubic polynomial in $[t_i,t_{i+1}]$, $i = 1,\ldots,n-1$

CS2: $\sigma \in C^2(t_1,t_n)$

It is well known that S_3 is a linear space of dimension $n+2$. So, if we want to use as interpolant a function $\sigma \in S_3$ we must impose two additional conditions in order to determine the $n+2$ free parameters. Among the most popular conditions we have:

I. Natural $\quad \sigma''(t_1) = \sigma''(t_n) = 0$

II. Periodic $\quad \sigma'(t_1) = \sigma'(t_n)$

$\quad \sigma''(t_1) = \sigma''(t_n)$

III. Hermite $\quad \sigma'(t_1) = y'_n$

$\quad \sigma''(t_n) = y'_n$.

Optimal convergence rates are obtained using conditions type II or III. In these lectures we will deal only with natural conditions. For a more detailed description of convergence rates the reader is referred to [9], [26] and the references therein.

Let $\mathcal{S}_3$ be the linear space of natural cubic splines, that is, the space of elements of S_3 satisfying condition I. It is an n-dimensional linear space and for any $y_1, \ldots, y_n \in \mathbb{R}$ there exist one and only one element in $\mathcal{S}_3$ such that

$$\sigma(t_i) = y_i \qquad i = 1,\ldots,n \ .$$

This new interpolant keeps some of the most important properties of polynomials:

1) The evaluation of σ at a given point t only requires multiplications and additions

2) The integral or derivative of σ is still a piecewise polynomial.

The advantage of this kind of interpolant is that we get "good" convergence rates even for functions being only continuous. (For details, see [9], [26]).

Of course we do not have an explicit formula like (1.1.2) and we will have to solve a linear system in order to find the n free parameters determining σ. Many methods are available, but we would like to point out one that is now able to give a generalization of splines to two and higher dimensions. It is the B-spline method. (Here B stands for basis) (cf. [5], [9], [26]).

The idea is to construct a basis for $\mathcal{S}_3$ with local support, that is, such that each element in the basis has the smallest possible support. Then the evaluation of σ at a point t needs the calculation of a linear combination with only a small number of nonzero terms.

More precisely, let $x_{-2} \le x_{-1} \le \cdots \le x_n \le \cdots \le x_{n+3}$ be the extended partition

$$x_{-2} = x_{-1} = x_0 = t_1$$

$$x_i = t_i \qquad i = 1,\ldots,n$$

$$x_{n+1} = x_{n+2} = x_{n+3} = t_n \ .$$

Now define the $n+2$ B-splines by the formula

$$B_i(t) = (-1)^m (x_{i+2} - x_{i-2})\ \delta_{x_{i-2},x_{i-1},x_i,x_{i+1},x_{i+2}} (t-x)_+^3 \ , \qquad (1.1.5)$$

$$t_1 \le t < t_n \ , \qquad i = 0,\ldots,n+1 \ ,$$

where $\delta_{\alpha_1,\ldots,\alpha_p} f$ is the divided difference of f with respect to $\alpha_1, \ldots, \alpha_p$. And the truncated power function $(t-x)_+^3$ is given by

$$(t-x)_+^3 = \begin{cases} 0 & t < x \\ (t-x)^3 & t \ge x \end{cases} \ . \qquad (1.1.6)$$

For $t = t_n$ we have

$$B_i(t_n) = \lim_{t \to t_n^-} B_i(t) \ . \qquad (1.1.7)$$

The set $\{B_0,\ldots,B_{n+1}\}$ is a basis for S_3 with the following interesting properties:

Theorem 1.1.1. The B-splines $B_0, \ldots, B_{n+1}$ defined by (1.1.5) - (1.1.7) satisfy

$$B_i(t) = 0 \qquad t \notin [x_{i-2},x_{i+2}) \qquad (1.1.8)$$

$$B_i(t) > 0 \qquad t \in (x_{i-2},x_{i+2}) \ . \qquad (1.1.9)$$

Moreover, for any $t \in [t_1, t_n]$ we have

$$\sum_{i=0}^{n+1} B_i(t) = 1 \quad . \tag{1.1.10}$$

Proof. See [26].

These remarkable properties imply among others that the linear system to be solved has a band matrix. More precisely, the matrix is tridiagonal in the cubic case. Another interesting fact extensively used for computational purposes is the recurrence relation which allows a very simple algorithm for the calculation of $B_i(t)$. For details see [26].

As we shall see in other sections, the first extension of spline techniques to higher dimensions has been given using tensor products of one-dimensional splines. The small support basis for that case is easily found using tensor products of B-splines. This kind of technique can be used in interpolation, smoothing or curve fitting, as we shall see later.

More recently there has been a lot of interest in the development of multi-dimensional B-splines. We are not going to discuss this approach during these lectures and the interested reader is referred to [7], [8] and the references therein.

Now we turn our attention to a different property of polynomial splines which allows a very natural extension to higher dimensions. We begin with a lemma.

Lemma 1.1.2. Let $g \in C^2[t_1, t_n]$, then for any $s \in \mathcal{S}_3$ we have:

$$\int_{t_1}^{t_n} g''(t)s''(t)\, dt = \sum_{i=1}^{n} (s'''(t_i^+) - s'''(t_i^-))\, g(t_i) \tag{1.1.11}$$

where

$$s'''(t_1^-) \equiv 0$$

$$s'''(t_n^-) \equiv 0 \ .$$

Proof. Integrating by parts, using the fact that s''' is piecewise constant and rearranging terms we get the desired result.

Now consider any function μ in $C^2[t_1,t_n]$ interpolating the data $y_1, \ldots, y_n$ at $t_1, \ldots, t_n$. And let σ be the unique element in S_3 interpolating the data. Then from (1.1.11) we have

$$\int_{t_1}^{t_n} (\mu''(t) - \sigma''(t))\, \sigma''(t)\, dt = 0 \ .$$

Or, if we extend σ to $[0,T]$ by

$$\sigma(t) = \begin{cases} \sigma(t_1) + (t - t_1)\sigma'(t_1) & t < t_1 \\ \sigma(t) & t_1 \le t \le t_n \\ \sigma(t_n) + (t - t_n)\sigma'(t_n) & t > t_n \end{cases}$$

we have

$$\int_0^T (\mu''(t) - \sigma''(t))\, \sigma''(t)\, dt = 0 \ . \tag{1.1.12}$$

More precisely, we have the following:

Theorem 1.1.3. For any $\mu \in H^2[0,T]$ interpolating $y_1, \ldots, y_n$ at $t_1, \ldots, t_n$, we have

$$(\mu-\sigma,\sigma) = 0$$

where σ is the natural cubic spline interpolating the data and

$$(\mu,\nu) = \int_0^T \mu''(t)\, \nu''(t)\, dt \quad . \tag{1.1.13}$$

Proof. See [19], [26].

From this theorem, usually called the First Integral Relation, we immediately get the following property that is usually taken as an equivalent definition of the natural cubic spline:

Theorem 1.1.4. The natural cubic spline σ interpolating $y_1, \ldots, y_n$ at $t_1, \ldots, y_n$ is the unique solution to the following minimization problem.

$$\underset{\mu \in I_y}{\text{Minimize}} \quad (\mu,\mu) \tag{1.1.14}$$

where

$$I_y = \{\mu \in H^2[0,T] \mid \mu(t_i) = y_i,\ i = 1,\ldots,n\} \quad .$$

Proof. From the First Integral Relation we get

$$\begin{aligned} (\mu,\mu) &= (\mu-\sigma+\sigma,\ \mu-\sigma+\sigma) \\ &= (\mu-\sigma,\ \mu-\sigma) + 2\cancelto{0}{(\sigma,\mu-\sigma)} + (\sigma,\sigma) \\ &= (\mu-\sigma,\ \mu-\sigma) + (\sigma,\sigma) \geq (\sigma,\sigma) \quad , \qquad \mu \in I_y \quad . \end{aligned}$$

This proves that σ minimizes $(\cdot,\cdot)$ over I_y. The uniqueness comes from the fact that the kernel of the semi-norm $(\cdot,\cdot)$ is P_1 the set of polynomials of degree one whose intersection with $I_0 = \{\mu \in H^2[0,T] \mid \mu(t_i) = 0\}$ is the element 0. This concludes the proof.

This variational definition of the natural cubic splines has lead to several generalizations in one and many dimensions. We are going to study some of these during these lectures. For additional details see [3], [7], [9], [13], [19], [26] and the references therein.

1.2. Smoothing Splines

Consider now the case where the data are noisy. This is the case in most practical problems. Assume the model

$$y_i = f(t_i) + \varepsilon_i \qquad i = 1,\dots,n \tag{1.2.1}$$

where f is the "smooth" unknown function to be approximated and the errors ε_i, $i = 1,\dots,n$ are assumed to be realizations of independent identically distributed normal random variables with

$$\begin{aligned} E[\varepsilon_i] &= 0 \\ E[\varepsilon_i^2] &= \nu^2 \qquad i = 1,\dots,n \\ E[\varepsilon_i \varepsilon_j] &= 0 \qquad i \neq j \ . \end{aligned}$$

To solve this problem Schoemberg introduced the smoothing spline. This spline is the unique solution to the following minimization problem.

$$\underset{\mu \in H^2[0,T]}{\text{Minimize}} \left\{ \lambda \int_0^T (\mu''(t))^2 dt + \frac{1}{n} \sum_{i=1}^{n} (\mu(t_i) - y_i)^2 \right\} \ . \tag{1.2.2}$$

Where $\lambda > 0$ is the smoothing parameter controlling the tradeoff between the smoothness of the solution measured by (μ,μ) and the approximation to the data measured by $\frac{1}{n} \sum_{i=1}^{n} (\mu(t_i) - y_i)^2$.

The solution to (1.2.2) is an element of S_3, that is, a natural cubic spline. Moreover, let Ω be the $n\times n$ symmetric semi-positive definite matrix such that

$$z^T \Omega z = \min_{\substack{\mu\in H^2[0,T] \\ \mu(t_i)=z_i \\ i=1,\dots,n}} (\mu,\mu) \qquad (1.2.3.)$$

This matrix is well defined because the transformation associating the interpolating cubic spline to the data $z_1, \dots, z_n$ is linear and $(\cdot,\cdot)$ is quadratic.

With this definition (1.2.2) can be restated as:

$$\underset{Z\in\mathbb{R}^n}{\text{Minimize}} \left\{ \underset{\substack{\mu\in H^2[0,T] \\ \mu(t_i)=z_i \\ i=1,\dots,n}}{\text{Minimize}} \left\{ \lambda(\mu,\mu) + \frac{1}{n}\sum_{i=1}^{n} (u(t_i) - y_i)^2 \right\} \right\} .$$

Or,

$$\underset{z\in\mathbb{R}^n}{\text{Minimize}} \left\{ \lambda\, z^T\Omega z + \frac{1}{n}\sum_{i=1}^{n} (y_i - z_i)^2 \right\} . \qquad (1.2.4)$$

And the solution $\hat{z}$ to this problem is then given by

$$\hat{z} = (I + n\lambda\Omega)^{-1} y . \qquad (1.2.5)$$

If we call $\sigma_{n,\lambda}$ to the smoothing spline then $\hat{z}$ is the vector of the values of $\sigma_{n,\lambda}$ at the knots

$$\hat{z}_i = \sigma_{n,\lambda}(t_i) \qquad i = 1,\dots,n . \qquad (1.2.6)$$

As we shall see later, (1.2.5) is not of practical use but it allows us to give a nice interpretation of the effect of this procedure on the data.

Ω being symmetric and positive semi-definite, its eigenvalues are positive real numbers. Moreover, we have (cf. [11]) for ρ_i, $i = 1,\dots,n$; the eigenvalues of Ω

$$0 = \rho_1 = \rho_2 < \rho_3 < \cdots < \rho_n \quad . \tag{1.2.7}$$

Also, for n sufficiently large and $i > 10$ (for practical purposes), we have (cf. [33]),

$$\rho_i \sim \alpha \frac{i^4}{n} \tag{1.2.8}$$

Let Q be the orthogonal matrix diagonalizing Ω, then

$$\Omega = Q \, \Sigma \, Q^* \tag{1.2.9}$$

where $\Sigma = \text{diag}\,(\rho_1, \rho_2, \dots, \rho_n)$. And

$$\hat{z} = Q(I + n\lambda\Sigma)^{-1}Q^*y$$

or

$$\tilde{z} = (I + n\lambda\Sigma)^{-1}\tilde{y} \tag{1.2.10}$$

with

$$\tilde{z} = Q^*\hat{z}$$

$$\tilde{y} = Q^*y \quad .$$

If we write (1.2.10) in terms of each component and use (1.2.8) we obtain

$$\tilde{z}_i \sim \frac{1}{1+\lambda\alpha i^4}\, y_i \quad . \tag{1.2.11}$$

Thus we are filtering the data using a low-pass Butterworth type filter with cutting frequency $(\lambda\alpha)^{-\frac{1}{4}}$ to be chosen.

The choice of λ from the data is a very complicated problem that has received a lot of attention during the last years. For details see [4], [17], [30], [31], [37], [38]. We are just going to describe two of them. The first one was proposed by Reinsch [24] and is closely related to the methods for choosing the regularization parameter in Tikhonov Regularization [38].

The idea is to solve the following equation for

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \sigma_{n,\lambda}(t_i)]^2 = \sigma^2 \tag{1.2.12}$$

or, equivalently

$$\frac{1}{n}\sum_{i=1}^{n}\frac{n^2\lambda^2\rho_i^2}{(1+n\lambda\rho_i)^2}\,\tilde{y}_i^2 = \sigma^2 \quad ,$$

an equation that can be solved using bisection or another nesting method.

As it can be seen, the method is indeed very simple and the idea is attractive. Nevertheless there is a lot of numerical and theoretical evidence showing that in most cases the solution obtained by this method gives a spline which tends to be too smooth, i.e., it tends to eliminate variations of the true function (cf. [44]).

The second method is called Generalized Cross-validation and was introduced in this form by Craven and Wabba [4]. The idea of the method is more difficult to introduce and the interested reader should go to their paper [4] for a detailed description. See also [17], [30], [31].

The GCV method reduces to the minimization over λ of the following loss function, usually referred to as the GCV function:

$$V(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \sigma_{n,\lambda}(t_i))^2}{\left(1 - \frac{1}{n}\mathrm{tr.}(A(\lambda))\right)^2} \tag{1.2.13}$$

where

$$A(\lambda) = (I + n\lambda\Omega)^{-1} \quad .$$

Special methods have been developed for the efficient computation of the minimizer of V. For details see [30], [31], [43].

2. Thin Plate Splines

In this section we introduce one of the most popular forms of bidimensional splines and give theoretical and computational details. We have chosen to start with these kind of splines because we think it is a natural generalization of the popular cubic spline even if the historic development was slightly different.

2.1. Theoretical Background

Let $t^i = (t_1^i, t_2^i)$ $i = 1,2,\ldots,n$ be n different points of the Euclidean plane $\mathbb{R}^2$ and let $z_1, \ldots, z_n$ be n real numbers: the observations. As in the one-dimensional case, we seek for a smooth function g, called the interpolant, such that

$$g(t^i) = z_i \qquad i = 1,2,\ldots,n \quad . \tag{2.1.1}$$

In order to define a function interpolating the data we have first to define a domain in $\mathbb{R}^2$ containing the data points. There are many possible choices and in some cases it will be defined by the problem itself, but while we have the choice we are going to take $\mathbb{R}^2$ itself as the domain of definition of the interpolant. This will allow us to use the powerful tool of Fourier Transform. For other choices, see [3], [29].

Following the ideas of Theorem 1.1.4, a direct generalization of the loss function to be minimized is given by

$$J(\mu) = \int_{\mathbb{R}^2} \left(\frac{\partial^2\mu}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2\mu}{\partial x\partial y}\right)^2 + \left(\frac{\partial^2\mu}{\partial y^2}\right)^2 \quad . \tag{2.1.2}$$

In order to set properly the minimization problem we have to introduce the following function space:

$$D^{-2}L^2(\mathbb{R}^2) = \{\mu : \mathbb{R}^2 \to \mathbb{R} \mid D^\alpha\mu \in L^2(\mathbb{R}^2),\ |\alpha| = 2\} \tag{2.1.3}$$

where $\alpha = (\alpha_1, \alpha_2)$, $|\alpha| = \alpha_1 + \alpha_2$.

$D^{-2}L^2$ is not equal to $H^2(\mathbb{R}^2)$. Indeed we have

$$H^2(\mathbb{R}^2) \subset D^{-2}L^2(\mathbb{R}^2) \quad . \tag{2.1.4}$$

$D^{-2}L^2$ is indeed a space of continuous functions (see [13], [14]) and $\frac{D^{-2}L^2}{P_1}$ is a Hilbert space with the norm

$$|\dot{\mu}| = [J(\mu)]^{\frac{1}{2}} \qquad \text{and} \qquad \mu \in \dot{\mu} \tag{2.1.5}$$

and P_1 is the space of polynomials of degree 1.

<u>Theorem 2.1.1</u>. If $\{t^i\}_{i=1,\dots,n}$ contains at least one P_1-unisolvent set, then there exists one and only one element σ belonging to $D^{-2}L^2(\mathbb{R}^2)$ such that

$$J(\sigma) = \min_{\substack{\mu \in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i) = z_i \\ i=1,\dots,n}} J(\mu) \quad . \tag{2.1.6}$$

σ is called the Thin Plate Spline interpolating $z_1, \ldots, z_n$ at $t^1, \ldots, t^n$.

<u>Proof</u>. $\frac{D^{-2}L^2}{P_1}$ is a Hilbert space continuously inbedded in $\frac{\mathbb{R}^{\mathbb{R}^2}}{P_1}$. The set

$$I_z = \{\mu \in D^{-2}L^2(\mathbb{R}^2) \mid \mu(t^i)=z_i,\ i=1,\ldots,n\} \tag{2.1.7}$$

is clearly nonvoid, and

$$\dot{I}_z = I_z + P_1$$

is closed in $\frac{D^{-2}L^2}{P_1}$ since $I_z + P_1$ is closed in $\frac{\mathbb{R}^{\mathbb{R}^2}}{P_1}$

$$\left(I_z + P_1 = \bigcap_{i=1}^{n} \{\delta_{t^i}(z_i)\} + P_1\right)$$

and the injection from $\frac{D^{-2}L^2}{P_1}$ into $\frac{\mathbb{R}^{\mathbb{R}^2}}{P_1}$ is continuous.

Then there exists a unique element $\dot{\sigma} \in \frac{D^{-2}L^2}{P_1}$ minimizing $|\dot{\mu}|^2$ over $\dot{I}_z$. The uniqueness of σ comes from the fact that any two elements in $\dot{\sigma}$ are different only in a polynomial of degree 1 and from the hypothesis we have $I_0 \cap P_1 = \{0\}$. //

In order to obtain the characterization of the solution we have to use convex analysis. Let

$$\chi_I(\mu) = \begin{cases} 0 & \mu \in I \\ +\infty & \text{otherwise} \end{cases} .$$

Our problem can then be put in the form

$$\underset{\mu \in D^{-2}L^2(\mathbb{R}^2)}{\text{minimize}} \{J(\mu) + \chi_I(\mu)\}$$

and the solution σ satisfies the inclusion equation

$$0 \in \partial(J + \chi_I)(\sigma) \quad . \tag{2.1.8}$$

Using now the subdifferential calculus (see [19]) we get

$$0 = \Delta^2\sigma - \sum_{i=1}^{n} \lambda_i \, \delta_{t^i} \tag{2.1.9}$$

where the λ_i are coefficients, δ_{t^i} are Dirac measures and $\Delta^2\sigma$ is the iterated Laplacian. Moreover, it is clear that the measure $\Delta^2\sigma$ must be orthogonal to P_1. Then

$$\sum_{i=1}^{n} \lambda_i \, p(t^i) = 0 \quad , \qquad p \in P_1 \tag{2.1.10}$$

Now we use the fact that (cf. [28])

$$\Delta^2(r^2 \log r) = \delta$$

to conclude that

$$\sigma(t) = \sum_{i=1}^{n} \lambda_i |t-t^i|^2 \log |t-t^i| + q(t) \tag{2.1.11}$$

where q is a polynomial of degree 1.

For a complete proof of these results, see [13].

2.2. The Computation of σ

From (2.1.11) - (2.1.10) we obtain the linear system to be solved in $\lambda_1, \ldots, \lambda_n, \alpha_1, \alpha_2, \alpha_3$:

$$\sum_{i=1}^{n} \lambda_i |t^j - t^i|^2 \log |t^j - t^i| + \alpha_1 t_1^j + \alpha_2 t_2^j + \alpha_3 = z_j \,, \qquad j = 1,\dots,n$$

$$\sum_{i=1}^{n} \lambda_i t_1^i = 0$$

$$\sum_{i=1}^{n} \lambda_i t_2^i = 0 \tag{2.2.1}$$

$$\sum_{i=1}^{n} \lambda_i = 0 \,.$$

Or, in matrix form

$$\begin{aligned} K\Lambda + E\alpha &= z \\ E^T\Lambda &= 0 \,, \end{aligned} \tag{2.2.2}$$

where K is an $n\times n$ symmetric matrix and E is $n\times 3$ matrix.

$$(K)_{ij} = |t^j - t^i|^2 \log |t^j - t^i| \,, \tag{2.2.3}$$

$$\begin{aligned} (E)_{i1} &= t_1^i \\ (E)i_2 &= t_2^i \\ (E)_{i3} &= 1 \,. \end{aligned} \tag{2.2.4}$$

Many problems arise. The first one is that if K and E are full matrices then the solution of this system requires $O(n^3)$ operations. This was not the case for one-dimensional splines where the matrix involved in the calculations had a band structure and the system can then be solved in $O(n)$ operations.

A second major practical problem is that (2.2.2) is ill-conditioned and a preconditioning will be necessary to solve it successfully. Finally, from (2.1.11) we see that the evaluation of σ at t requires to use all the coefficients. This tends to be unstable due to the alternation of signs of the coefficients specially when n is large (depending upon the computer).

The ill conditioning of the system can be virtually eliminated using the technique we describe now.

Let Q, R be the Q-R decomposition of E. That is, let Q be an $n \times n$ orthogonal matrix and R be a 3×3 upper triangular nonsingular matrix, such that

$$Q^*E = \begin{vmatrix} R \\ 0 \end{vmatrix} . \tag{2.2.5}$$

And let us partition Q in the following way

$$Q^* = \begin{bmatrix} Q_1^* \\ Q_2^* \end{bmatrix} \tag{2.2.6}$$

where Q_1^* is $3 \times n$ and Q_2^* is $(n-3) \times n$.

Let

$$\tilde{\Lambda} = Q^*\Lambda$$

$$= \begin{bmatrix} Q_1^*\Lambda \\ Q_2^*\Lambda \end{bmatrix} = \begin{bmatrix} \tilde{\Lambda}_1 \\ \tilde{\Lambda}_2 \end{bmatrix} \tag{2.2.7}$$

where $\tilde{\Lambda}_1 \in \mathbb{R}^3$ and $\tilde{\Lambda}_2 \in \mathbb{R}^{n-3}$.

The second equation now becomes:

$$E^T\Lambda = 0$$

$$E^T Q \tilde{\Lambda} = 0$$

$$[R^T|0]\tilde{\Lambda} = 0$$

$$R^T\tilde{\Lambda}_1 = 0$$

$$\tilde{\Lambda}_1 = 0 \quad . \qquad (2.2.8)$$

The last equation is obtained using the nonsingularity of R. R is non-nonsingular since $t^1, \ldots, t^n$ contains at least a P_1-unisolvent set.

We now use the first equation

$$KQ\tilde{\Lambda} + E\alpha = z$$

$$[KQ_1 \mid KQ_2] \begin{vmatrix} \tilde{\Lambda}_1 \\ \tilde{\Lambda}_2 \end{vmatrix} + Q\begin{vmatrix} R \\ 0 \end{vmatrix}\alpha = z$$

$$KQ_2\tilde{\Lambda}_2 + Q_1 R\alpha = z \quad . \qquad (2.2.9)$$

Multiplying now both sides by Q_2^* we obtain

$$Q_2^* K Q_2\tilde{\Lambda}_2 + Q_2^* Q_1 R\alpha = Q_2^* z$$

or

$$Q_2^* K Q_2\tilde{\Lambda}_2 = Q_2^* z$$

since

$$Q_2^* Q_1 = 0 \quad .$$

Or, finally

$$A\tilde{\Lambda}_2 = \tilde{z} \qquad (2.2.10)$$

where

$$A = Q_2^* K Q_2$$

$$\tilde{z} = Q_2^* z \quad .$$

α can be easily obtained from (2.2.9)

$$Q_1 R\alpha = z - KQ_2\tilde{\Lambda}_2$$

or

$$Q_1^* Q_1 R\alpha = Q_1^* z - Q_1^* K Q_2 \tilde{\Lambda}_2$$

$$\alpha = R^{-1}[Q_1^* z - Q_1^* K Q_2 \tilde{\Lambda}_2] \quad . \tag{2.2.11}$$

For an efficient computer evaluation of the products Q_1^*z, Q_2^*z, $Q_2^* K Q_2$, let us simply observe that

$$\begin{bmatrix} Q_1^* z \\ Q_2^* z \end{bmatrix} = Q^* z$$

$$Q^* K Q = \begin{bmatrix} Q_1^* K Q_1 & Q_1^* K Q_2 \\ Q_2^* K Q_1 & Q_2^* K Q_2 \end{bmatrix} \quad .$$

Now the computation of Q^*z and Q^*KQ is easily done since Q, Q^* are obtained by a product of Householder transformations that is available for the computation of such products when using a package like LINPACK (see [12]).

In order to solve (2.2.10) we first prove the following.

Lemma 2.2.1. The matrix $Q_2^* K Q_2$ is positive definite.

Proof. The columns of Q_1 generate E and the columns of Q_2 generate $E^{\perp}$, the orthogonal subspace of E. Then Q_2 represents the orthogonal projection onto $E^{\perp}$ written in terms of the basis given by the columns of Q_2. Thus

$$\mu = Q_2 x \in E^{\perp} , \qquad x \in \mathbb{R}^{n-3} .$$

On the other hand the function

$$\omega(t) = \sum_{i=1}^{n} \mu_i |t-t^i|^2 \log |t-t^i| \tag{2.2.12}$$

with $\mu \in E^{\perp}$ being a spline function and

$$0 < J(\omega) = \mu^T K \mu \tag{2.2.13}$$

since from (2.1.9)

$$J(\omega) = \sum_{i=1}^{n} \mu_i \, \omega(t^i) ,$$

and $J(\omega)$ cannot be zero since ω is not a polynomial of degree 1. //

Now the system (2.2.10) can be solved using Cholesky's factorization.

As an interesting fact we might notice that A does not depend on the data itself but only on the data points. So if many data taken at the same data points are to be processed, the storage of the Cholesky's factorization of A allows us to save a lot of computer time.

For a complete description of this method and a discussion of its numerical properties, see [21], [22], [43].

As we have already said, the computation of thin plate splines can be very expensive if a large number of points is to be used. In the next section we develop a practical method to avoid this problem in some cases. Here we must also mention that the actual work towards the development of multidimensional B-splines could be of great help when dealing with a large number of points (cf. [7], [8], [26]).

2.3. A Practical Method for a Large Number of Points

In this section we assume that the number of data points is large and well distributed within a rectangle $[a,b]\,[c,d] \subset \mathbb{R}^2$. The idea (cf. [21], [22]) is to divide R, the given rectangle, into several smaller rectangles overlapping each other and containing a reasonable number of points. We then calculate a thin plate spline in each piece and stick the pieces together using a partition of unity.

More precisely, let $a = a_0 < a_1 < \cdots < a_m < b$, $a < b_0 < b_1 < \cdots < b_m = b$ and $c = c_0 < c_1 < \cdots < c_k < c$, $d < d_0 < d_1 < \cdots < d_k = d$ be partitions of $[a,b]$ and $[c,d]$ respectively such that

$$\begin{aligned} a_i &< b_i \qquad & i = 0,\ldots,m \\ c_i &< d_i & i = 0,\ldots,k \\ a_{i+1} &< b_i & i = 0,\ldots,m-1 \\ c_{i+1} &< d_i & i = 0,\ldots,k \quad . \end{aligned} \tag{2.3.1}$$

These properties imply that

$$(a_i,b_i) \cap (a_{i+1},b_{i+1}) \neq \emptyset$$

$$(c_i,d_i) \cap (c_{i+1},d_{i+1}) \neq \emptyset$$

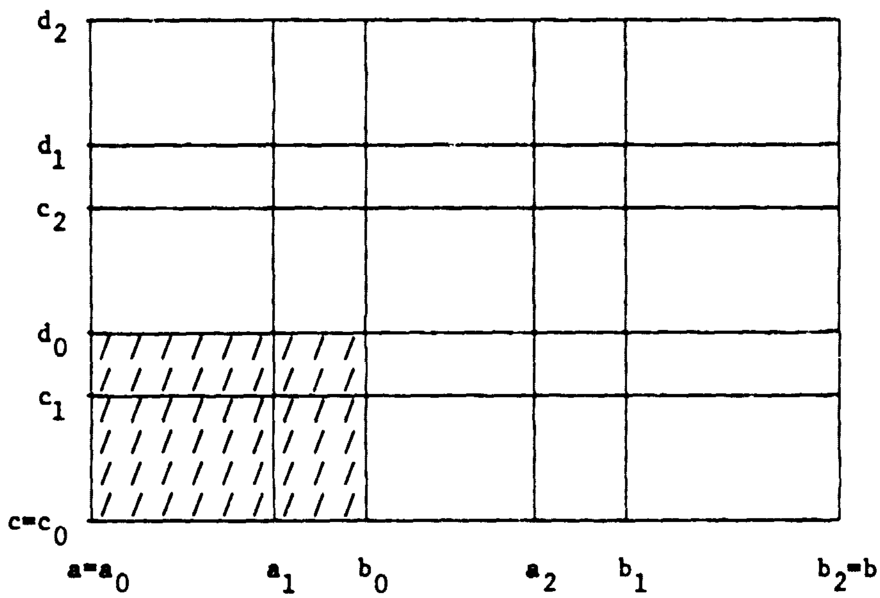
$$\bigcup_{i=1}^{m} [a_i,b_i] = [a,b]$$

$$\bigcup_{j=0}^{k} [c_k,d_k] = [c,d]$$

and

$$\bigcup_{i=1}^{m} \bigcup_{j=0}^{k} [a_i,b_i] \times [c_j,d_j] = [a,b] \times [c,d] \quad .$$

The situation is illustrated in the following figure



We now assume that the number of data points included in $R_{ij} = [a_i,b_i] \times [c_j,d_j]$ is "reasonable." This means that: 1) The computer time spent in the computation of a thin plate spline using that number of points is reasonable according to the particular problem. 2) The number of points is sufficiently large to represent the behavior of the function in that region. 3) The number of points in the intersection of the regions is enough to guarantee a smooth transfer from one region to another.

Of course all these conditions are extremely subjective and the user has to be experienced to know how to apply these conditions to his particular problem.

Once the subdivision of R in R_{ij} has been determined, let σ_{ij} be the thin plate spline interpolating the data on the points belonging to R_{ij}. That is,

$$J(\sigma_{ij}) = \min_{\substack{\mu \in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^k)=z_k \\ k \in I_{ij}}} J(\mu) \tag{2.3.2}$$

where

$$I_{ij} = \{k \in \{1,\ldots,n\} \mid t^k \in R_{ij} = [a_i,b_i]\times[c_j,d_j]\} \quad .$$

Thus σ_{ij} is going to be a "good" approximation of the unknown function on R_{ij}. In order to build an interpolant over R, we now "stick" the pieces together using a partition of unity on the intersection of the rectangles.

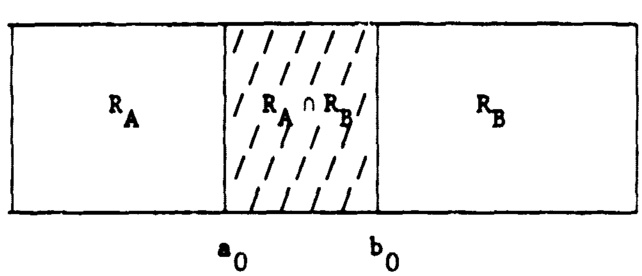
Let $\omega: \mathbb{R} \to \mathbb{R}$ be defined by

$$\omega(x) = \begin{cases} (x-1)^2(2x+1) & x \in [0,1] \\ 0 & x > 1 \\ 1 & x < 0 \end{cases} \quad . \tag{2.3.3}$$

Then $\omega \in C^1(\mathbb{R})$ and

$$\omega(x) + (1-\omega(x)) = 1 \qquad x \in \mathbb{R} \quad . \tag{2.3.4}$$

<u>Theorem 2.3.1</u>. Let σ_A, σ_B be thin plate splines interpolating the data over the rectangles R_A and R_B respectively.



Let σ_{AB} be defined by

$$\sigma_{AB}(x,y) = \begin{cases} \sigma_A(x,y) & (x,y) \in R_A \setminus R_B \\ \rho(x)\sigma_A(x,y) + (1-\rho(x))\sigma_B(x,y) & (x,y) \in R_A \cap R_B \\ \sigma_B(x,y) & (x,y) \in R_B \setminus R_A \end{cases}$$

where

$$\rho(x) = \omega\left(\frac{x - a_0}{b_0 - a_0}\right) .$$

Then σ_{AB} interpolates the data on $R_A \cup R_B$ and has continuous first partial derivatives.

<u>Proof</u>. It is clear that σ_{AB} interpolates the data on $R_A \cup R_B - R_A \cap R_B$. Let $t^k \in R_A \cap R_B$, then

$$\sigma_{AB}(t^k) = \rho(x^k)\sigma_A(t^k) + (1 - \rho(x^k))\sigma_B(t^k)$$

$$= \rho(x^k) z_k + (1 - \rho(x^k)) z_k = z_k$$

where $t^k = (x^k, y^k)$.

We only need to prove the continuity of σ_{AB} and $\frac{\partial \sigma_{AB}}{\partial x}$ along $x = a_0$, $x = b_0$. Both cases being similar we prove it along $x = a_0$. We have

$$\lim_{x \to a_0^+} \sigma_{AB}(x,y) = \lim_{x \to a_0^+} \{\rho(x)\sigma_A(x,y) + (1-\rho(x))\sigma_B(x,y)\}$$

$$= \sigma_A(a_0, y) .$$

Thus σ_{AB} is continuous across the line $x = a_0$. For $\frac{\partial \sigma_{AB}}{\partial x}$ we have

$$\frac{\partial\sigma_{AB}}{\partial x}(x,y) = \rho'(x)\sigma_A(x,y) + \rho(x)\frac{\partial\sigma_A(x,y)}{\partial x} + (1-\rho(x))\frac{\partial\sigma_B(x,y)}{\partial x} - \rho'(x)\frac{\partial\sigma_B(x,y)}{\partial x}$$

$$\lim_{x\to a_0^+}\frac{\partial\sigma_{AB}}{\partial x}(x,y) = \frac{\partial\sigma_A(a_0,y)}{\partial x}$$

and the result follows. //

Using this method we can stick together all the pieces on each horizontal strip and then proceed in the same way to stick the strips together. The final result is the following

$$G(t) = \begin{cases} \sigma_{ij}(t) & t \in R_{ij} \setminus [R_{ij-1} \cup R_{ij+1} \cup R_{i+1j} \cup R_{i-1j}] \\ \omega\left(\frac{x-a_{i+1}}{b_i-a_{i+1}}\right)\sigma_{ij}(t) + \left(1-\omega\left(\frac{x-a_{i+1}}{b_i-a_{i+1}}\right)\right)\sigma_{i+1j}(t) & \\ & t \in R_{ij} \cap R_{i+1j} \setminus (R_{ij+1} \cup R_{ij-1}) \\ \omega\left(\frac{y-c_{j+1}}{d_j-c_{j+1}}\right)\sigma_{ij}(t) + \left(1-\omega\left(\frac{y-c_{j+1}}{d_j-c_{j+1}}\right)\right)\sigma_{ij+1}(t) & \\ & t \in R_{ij} \cap R_{ij+1} \setminus (R_{i+1j} \cup R_{i-1j} \\ \omega\left(\frac{x-a_{i+1}}{b_i-a_{i+1}}\right)\left[\omega\left(\frac{y-c_{j+1}}{d_j-c_{j+1}}\right)\sigma_{ij}(t) + \left(1-\omega\left(\frac{y-c_{j+1}}{d_j-c_{j+1}}\right)\right)\sigma_{i,j+1}(t)\right] & \\ \quad + \left(1-\omega\left(\frac{x-a_{i+1}}{b_i-a_{i+1}}\right)\right)\left[\omega\left(\frac{y-c_{j+1}}{d_j-c_{j+1}}\right)\sigma_{i+1j}(t) + \left(1-\omega\left(\frac{y-c_{j+1}}{d_j-c_{j+1}}\right)\right)\sigma_{i+1j+1}(t)\right] & \\ & t \in R_{ij} \cap R_{i+1j} \cap R_{ij+1} \cap R_{i+1j+1}\,. \end{cases} \tag{2.3.4}$$

It has been proved by Duchon [14] that this interpolant keeps the nice convergence properties of a thin plate spline. On the other hand our numerical experience shows that it is a very appropriate method to deal with a large number of points.

Remark. In the above formula we have made the convention

$$\sigma_{0,j} = \sigma_{m+1,j} = \sigma_{i,0} = \sigma_{i,k+1} \equiv 0 \ .$$

3. A General Framework

The variational properties of natural polynomial splines in one dimension lead Atteia [3] to give a general variational formulation for spline functions. This idea is to consider three Hilbert spaces X, Y, Z and two continuous linear operators T, A from X onto Y and X onto Z respectively, then given $z \in Z$ a general spline of interpolation is the solution to

$$\min_{\substack{\mu \in X \\ A(\mu)=z}} \|T\mu\|_Y^2 \ .$$

All the known cases of splines satisfying variational properties can be found using this general approach. For details see [19]. Nevertheless, in some cases it is somewhat difficult or not natural to use some Hilbert space to make the splines fit into this model. For that reason Duchon [14] introduced the semi-Hilbert space model where he works with semi-reproducing kernels. In this section we are going to give a slightly different formulation using reproducing kernels because of its immediate stochastic interpretation.

3.1. Reproducing Kernels and Spline Functions

Let H be a repfoducing kernel Hilbert space (see [2], [27]) that is, a Hilbert space where the evaluation functionals are continuous. H must be a subset of $\mathbb{R}^X$, the space of functions from X into $\mathbb{R}$.

$$H \subset \mathbb{R}^X \quad . \tag{3.1.1}$$

Following Aronzajn [2] this implies the existence of a function K, the reproducing kernel, such that

$$\begin{array}{lll} \text{i)} & K(\cdot,x) \in H & \text{any } x \in X \\ \text{ii)} & K(x,y) = K(y,x) & \\ \text{iii)} & f(x) = \langle f,K(\cdot,x)\rangle & \text{any } f \in H \end{array} \tag{3.1.2}$$

where $\langle \cdot,\cdot \rangle$ denotes the scalar product in H.

Let $\lambda_1, \ldots, \lambda_n$ be continuous linear functionals from H onto $\mathbb{R}$. And let $z \in \mathbb{R}^n$ be an arbitrary vector of real numbers. We define the spline interpolating $z_1, \ldots, z_n$ with respect to the functionals $\lambda_1, \ldots, \lambda_n$ as the unique solution to

$$\min_{\substack{\mu \in H \\ \lambda_i(u)=z_i \\ i=1,\ldots,n}} \|\mu\| \quad . \tag{3.1.3}$$

The existence and uniqueness of $\hat{\sigma}$ the solution to (3.1.3) is a result of the projection theorem in a Hilbert space and the fact that $\{\mu \in H \mid \lambda_i(u)=z_i\}$ is a closed linear manifold in H.

The special interest of working with a reproducing kernel Hilbert space is based in the following result.

<u>Theorem 3.1.1</u>. The unique solution to (3.1.3) is given by

$$\hat{\sigma}(x) = \sum_{i=1}^{n} \alpha_i \, h_i(x) \tag{3.1.4}$$

where

$$h_i(x) = \lambda_i(K(\cdot,x)) \tag{3.1.5}$$

and $\alpha_1, \ldots, \alpha_n$ satisfy the linear system

$$\sum_{j=1}^{n} \alpha_j \, \lambda_i(h_j) = z_i \quad , \qquad i = 1,\ldots,n \quad . \tag{3.1.6}$$

<u>Proof</u>. Let $A\colon H \to \mathbb{R}^n$ be the linear transformation

$$A(\mu) = (\lambda_1(\mu), \ldots, \lambda_n(\mu))^T \quad .$$

It is clear that $\hat{\sigma}$, the solution to (3.1.3), is the orthogonal projection of 0 onto the linear closed manifold

$$M = \{\mu \in H \mid \lambda_i(u)=z_i, \ i=1,\ldots,n\} \quad .$$

exists and is unique if and only if M is non-void. Moreover, we have the orthogonality of the projection:

$$\langle 0-\hat{\sigma}, \mu \rangle = 0 \qquad \text{if} \quad A(\mu) = 0 \quad .$$

In other words, $\hat{\sigma} \in N(A)^{\perp} = R(A')$. But

$$A'(x_1,\ldots,x_n) = \sum_{i=1}^{n} x_i h_i$$

where $h_i(t) = \lambda_i(K(\cdot,t))$ is the representation of λ_i guaranteed by the

Ritz theorem. Then there exist $\alpha_1, \ldots, \alpha_n$ such that

$$\hat{\sigma} = \sum_{i=1}^{n} \alpha_i h_i \quad .$$

But $\hat{\sigma}$ also belongs to M, hence

$$\lambda_i(\hat{\sigma}) = \sum_{j=1}^{n} \alpha_j \, \lambda_i(h_j) = z_j \quad .$$

This completes the proof. //

As an example of use of this theorem we could try to get the thin plate spline.

Of course here the functionals $\lambda_1, \ldots, \lambda_n$ are given by

$$\lambda_i(\mu) = \mu(t^i) \qquad i = 1,\ldots,n \quad . \tag{3.1.7}$$

Let us assume that $\{t^1, t^2, t^3\}$ is a P_1-unisolvent set. And let $\{p_1, p_2, p_3\}$ be a basis of P_1 satisfying

$$\sum_{k=1}^{3} p_i(t^k) \, p_j(t^k) = \delta_{ij} \quad . \tag{3.1.8}$$

Then it can be shown that (cf. [43]) $D^{-2}L^2(\mathbb{R}^2)$ is a Hilbert space with norm

$$||\mu||^2 = \sum_{k=1}^{3} [u(t^k)]^2 + \int_{\mathbb{R}^2} \left(\frac{\partial^2 \mu}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 \mu}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 \mu}{\partial y^2}\right)^2 \tag{3.1.9}$$

and reproducing kernel

$$K(t,s) = \frac{1}{8\pi} |t-s|^2 \log |t-s| - \frac{1}{8\pi} \sum_{k=1}^{3} p_k(t) \; |s-t^k|^2 \log |s-t^k|$$

$$- \frac{1}{8\pi} \sum_{k=1}^{3} p_k(s) \; |t-t^k|^2 \log |t-t^k| \tag{3.1.10}$$

$$+ \frac{1}{8\pi} \sum_{i,j=1}^{3} p_i(t) p_j(s) \; |t^j-t^i|^2 \log |t^j-t^i| + \sum_{k=1}^{3} p_k(t) \; p_k(s) \quad .$$

Now using the theorem, the thin plate splines can be found solving the linear system

$$\sum_{j=1}^{n} \alpha_j \, K(t,t^j) = z_j \qquad j = 1,\dots,n \quad . \tag{3.1.11}$$

3.2. The Stochastic Approach

Let ω be a random process indexed by X with covariance

$$E[\omega_s \omega_t] = K(s,t) \tag{3.2.1}$$

and zero mean.

Let $L^2(\omega)$ be the Hilbert space spanned by ω_x, $x \in X$ and their limits in quadratic mean. $L^2(\omega)$ has the inner product

$$[\mu,\nu] = E[\mu\nu] \quad . \tag{3.2.2}$$

As in the last section, we denote by H the Hilbert space with the reproducing kernel K.

It is a well-known result (cf. [10]) that H and $L^2(\omega)$ are isomorphic with the correspondence

$$f(x) = E[\omega_x \mu] \tag{3.2.3}$$

$$f \in H \quad \leftrightarrow \quad \mu \in L^2(\omega) \quad .$$

Let $\lambda_1, \ldots, \lambda_n$ be the linear functional of thepreceding section. It is known that $\lambda_1, \ldots, \lambda_n$ can be defined on ω using the isometry. Thus $\lambda_i(\omega)$ is going to be an element of $L^2(\omega)$ defined by the equation

$$h_i(x) = E[\omega_x \lambda_i(\omega)] \quad . \tag{3.2.4}$$

We consider now the problem of finding thebest linear estimator for ω_x given $\lambda_1(\omega), \ldots, \lambda_n(\omega)$. That is, we want to find

$$\tilde{\omega}_x = E[\omega_x \mid \lambda_1(\omega), \ldots, \lambda_n(\omega)] \quad . \tag{3.2.5}$$

If we assume that the process is Gaussian, we can find $\tilde{\omega}_x$ as the least squares predictor of ω_x, i.e., as the projection in $L^2(\omega)$ of ω_x onto the linear space spanned by $\lambda_1(\omega), \ldots, \lambda_n(\omega)$.

Thus, $\tilde{\omega}_x$ is the solution to

$$\underset{(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n}{\text{minimize}} \left\| \sum_{i=1}^{n} \alpha_i \lambda_i(\omega) - \omega_x \right\|^2_{L^2(\omega)} \tag{3.2.6}$$

and

$$\tilde{\omega}_x = \sum_{i=1}^{n} \tilde{\alpha}_i \lambda_i(\omega) \quad . \tag{3.2.7}$$

In order to find $\tilde{\alpha}_i$, $i = 1, \ldots, n$, we have to solve the system of linear equations given by the orthogonality conditions

$$E\left[\left(\sum_{i=1}^{n} \tilde{\alpha}_i \lambda_i(\omega) - \omega_x\right) \lambda_j(\omega)\right] = 0 \qquad j = 1, \ldots, n \quad , \tag{3.2.8}$$

or, equivalently,

$$\sum_{i=1}^{n} \tilde{\alpha}_i \, E[\lambda_i(\omega) \lambda_j(\omega)] = E[\omega_x \lambda_j(\omega)]$$

and using the isometry again

$$\sum_{i=1}^{n} \tilde{\alpha}_i \langle h_i, h_j \rangle_H = h_j(x) \quad .$$

Finally, using the fact that h_j is the representer of λ_j we get

$$\sum_{i=1}^{n} \tilde{\alpha}_i \lambda_j(h_i) = h_j(x) \quad . \tag{3.2.9}$$

If we now assume that $z_1, \ldots, z_n$ are realizations of the random variables $\lambda_1(\omega), \ldots, \lambda_n(\omega)$ we get

$$\hat{\omega}_x = \sum_{i=1}^{n} \hat{\alpha}_i z_i \tag{3.2.10}$$

where

$$\hat{\alpha} = \Sigma^{-1} (h_1(x), \ldots, h_n(x))^T$$

in other words

$$\hat{\omega}_x = z^T \Sigma^{-1} (h_1(x), \ldots, h_n(x))^T \quad . \tag{3.2.11}$$

If we now go back to Theorem 3.1.1 we see that

$$\hat{\sigma}(x) = \hat{\omega}_x \quad .$$

Theorem 3.2.1. Let ω be a random Gaussian process indexed by X with zero mean and covariance function $K(\cdot,\cdot)$. Then

$$\hat{\sigma}(x) = E[\omega_x \mid \lambda_i(\omega)=z_i, \ i=1,\ldots,n] \quad . \tag{3.2.12}$$

For further results see [ ], [ ], [ ], [ ].

4. Constrained Interpolation

In the preceding sections we have considered the problem of interpolation of an unknown function given its values at n are generally distributed points of the Euclidean space $\mathbb{R}^2$. The hypothesis made on the function is generally one of the kind: "f is smooth." This meaning, most times $f \in D^{-2}L^2(\mathbb{R}^2)$ or something similar, but in all cases we have assumed that f belongs to some linear space and have used as interpolant the minimizer of a semi-norm over the linear manifold of interpolants belonging to that linear space. As a result, the interpolant is linear as a function of the data and the algorithms necessary to compute it require the solution of matrix problems. Even if this is sufficient for many problems, there are others where some nonlinear constraint is naturally posed, as in the case of data coming from a positive function. More generally, we could assume that some property of the function is known in the form $f \in C$ where C is a convex closed set of $D^{-2}L^2(\mathbb{R}^2)$. The problem is to find a "good" interpolant satisfying the nonlinear constraint $f \in C$.

This problem has recently received much attention in one and two dimensions. The interested reader should see the literature [18], [32], [35], [36], and the references therein.

4.1. The Positive Thin Plate Spline

For these lectures we would only be interested in positive thin plate splines (cf. [35], [36]). Let $z_1, z_2, \ldots, z_n$ be positive real numbers and let C be a compact subset of $\mathbb{R}^2$. We define the Positive Thin Plate Spline as the unique solution to the minimization problem (cf. Th. 4.1.1).

$$\begin{array}{l} \underset{\substack{\mu\in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^j)=z_j,\ j=1,\dots,n \\ \mu(t)\geq 0,\ t\in C}}{\text{minimize}} \quad J(\mu) \end{array} \tag{4.1.1}$$

where, as in the preceding sections J is defined by

$$J(\mu) = \int_{\mathbb{R}^2} \left[\frac{\partial^2\mu}{\partial x^2}\right]^2 + 2\left[\frac{\partial^2\mu}{\partial x\partial y}\right]^2 + \left[\frac{\partial^2\mu}{\partial y^2}\right]^2 .$$

We first give an existence and uniqueness theorem.

Theorem 4.1.1. Let C be compact and assume that $\{t^1, t^2, \dots, t^n\}$ contains at least a P_1-unisolvent set. Then there exists a unique solution of (4.1.1).

Proof. Let P_1 be the set of polynomials of degree 1, then $H = \frac{D^{-2}L^2(\mathbb{R}^2)}{P_1}$ is a Hilbert space continuously imbedded in $\frac{\mathbb{R}^{\mathbb{R}^2}}{P_1}$. The set

$$S = \{\mu\in D^{-2}L^2(\mathbb{R}^2) \mid \mu(t^i)=z_i,\ i=1,\dots,n;\ \mu(t)\geq 0,\ t\in C\}$$

is clearly non-void and

$$\dot{S} = S + P_1 \subseteq H$$

is closed since $S + P_1$ is closed in $\frac{\mathbb{R}^{\mathbb{R}^2}}{P_1}$.

$$S + P_1 = \left\{\bigcap_{i=1}^{n} \delta_{t^i}^{-1}(z_i) \bigcap_{t\in C} \delta_t^{-1}([0,\infty))\right\} + P_1$$

and H is continuously imbedded in $\frac{\mathbb{R}^{\mathbb{R}^2}}{P_1}$.

Then there exists a unique element $\dot{\sigma}\in H$ minimizing $J(\dot{\mu})$ over $\dot{S}$ ($J(\dot{\mu})$ is a norm in H.) The uniqueness of σ comes from the fact that

$$P_1 \cap \{\mu \in D^{-2}L^2(\mathbb{R}^2) \mid \mu(t^i)=0,\ i=1,\dots,n\} = \{0\} \quad ,$$

since $\{t^1, \dots, t^n\}$ contains a P_1-unisolvent set. //

It is also possible to give a general characterization of the solution σ. We give this characterization in Th. 4.1.2 but we omit the proof. The interested reader is referred to [35].

Theorem 4.1.2. Let σ be the unique solution of (4.1.1) and define

$$A = \{t \in C \mid \sigma(t)=0\} \quad . \tag{4.1.2}$$

A is a closed set. Then there exist a positive measure ν with support $(\nu) \subset A$ and n real numbers $\lambda_1, \dots, \lambda_n$ such that

$$\sigma = -\sum_{i=1}^{n} \lambda_i \, \delta_{t^i} * K_2 - \nu * K_2 + p \tag{4.1.3}$$

where $p \in P_1$ and K_2 is the elementary solution of the biharmonic operator

$$K_2(r) = \frac{1}{8\pi} P_f(r^2 \log r) \quad . \tag{4.1.4}$$

Here $*$ denotes the product of convolution. //

This last result is only of theoretical importance since it has not been possible to use it in a numerical algorithm allowing σ to be found. In the next section we present a numerical procedure allowing the solution to be found. For convergence properties of constrained splines see [32], [35].

4.2. The Dual Algorithm

Our aim in this section is to present a special version of the general algorithm by Laurent to solve constrained problems by dual iteration (cf. [20]).

We also mention the special numerical methods used to compute the k^{th} iterate.

The idea of the algorithm is to construct a sequence of half-spaces D_k and solve in each iteration the constrained problem.

$$(P_k) \qquad \min_{\substack{\mu \in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i)=z_i,\ i=1,\dots,n \\ \mu \in D_k}} J(\mu) \quad . \tag{4.2.1}$$

The sequence $\{\sigma_k\}$ is then shown to satisfy the following properties (cf. [20])

i) $J(\sigma_k) \le J(\sigma_{k+1}) \le J(\sigma)$

ii) $D_k \supset \{\mu \in D^{-2}L^2(\mathbb{R}^2) \mid \mu(t) \ge 0,\ t \in C\}$

iii) $\lim_{k\to\infty} \sigma_k = \sigma$

where the convergence can be taken over $H^2(C)$.

The starting point of the sequence is the unconstrained thin plate spline. That is, the solution of

$$J(\sigma_0) = \min_{\substack{\mu \in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i)=z_i,\ i=1,\dots,n}} J(\mu) \quad . \tag{4.2.2}$$

<u>Step 0</u>. Let $b_0 \in \mathbb{R}^2$ be a solution of

$$\sigma_0(b_0) = \min_{t \in C} \sigma(t) \quad . \tag{4.2.3}$$

If $\sigma_0(b_0) \geq 0$ then $\sigma = \sigma_0$ and the algorithm stops. Otherwise set $k = 0$ and proceed to step 1.

Step 1. Let

$$D_1 = \{\mu \in D^{-2}L^2(\mathbb{R}^2) \mid \mu(b_0) \geq 0\} \tag{4.2.4}$$

and let σ_1 be the unique solution of

$$\underset{\substack{\mu \in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i)=z_i,\ i=1,\ldots,n \\ \mu \in D_1}}{\text{minimize}} \quad J(\mu) \quad . \tag{4.2.5}$$

And set b_1 as a solution to

$$\sigma_1(b_1) = \min_{t \in C} \sigma_1(t) \quad . \tag{4.2.6}$$

Step 2. Let $k \leftarrow k+1$.

If we are lucky enough $\sigma_k(b_k) \geq 0$ and $\sigma = \sigma_k$. Stop.

Otherwise, let

$$\tilde{D}_k = D_k \cap \{\mu \in D^{-2}L^2(\mathbb{R}^2) \mid \mu(b_k) \geq 0\} \tag{4.2.7}$$

and let σ_{k+1} be the solution to

$$\underset{\substack{\mu \in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i)=z_i,\ i=1,\ldots,n \\ \mu \in \tilde{D}_k}}{\text{minimize}} \quad J(\mu) \quad . \tag{4.2.8}$$

Using now the Kuhn-Tucker conditions we know that there exist positive constants ν_j^k, $j = 1,\ldots,n_k$ such that if we set

$$D_{k+1} = \left\{ \mu \in D^{-2}L^2(\mathbb{R}^2) \;\middle|\; \sum_{j=1}^{n_k} \nu_j^k \, \mu(b_j^k) = 0 \right\} \tag{4.2.9}$$

we have

$$J(\sigma_{k+1}) = \min_{\substack{\mu \in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i) = z_i,\ i=1,\dots,n \\ \mu \in D_{k+1}}} J(\mu) \quad . \tag{4.2.10}$$

Now define b_{k+1} as a solution to

$$\sigma_{k+1}(b_{k+1}) = \min_{t \in C} \sigma_{k+1}(t) \tag{4.2.11}$$

and go to step k.

The preceding calculations are repeated until a convergence criterion is satisfied. Generally this criterion is given in the form

$$\sigma_{k+1}(b_{k+1}) \geq -\varepsilon \tag{4.2.12}$$

where $\varepsilon > 0$ is a tolerance given by the user.

Remarks

(1) The computation of σ_1 is not as hard as it seems to be since

$$\sigma_1(b_0) = 0 \quad . \tag{4.2.13}$$

Otherwise the constraint $\mu \in D_1$ would not be active and the solution might give $\sigma_1(b_0) > 0$ which is impossible since in that case $\sigma_0 = \sigma_1$.

Thus σ_1 is indeed the solution to the simple interpolation problem

$$\begin{array}{l} \text{minimize} \quad J(\mu) \\ \mu\in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i)=z_i,\ i=1,\dots,n \\ \mu(b_0)=0 \end{array} \qquad . \tag{4.2.14}$$

And σ_1 is of the form

$$\sigma_1(t) = \sum_{i=1}^{n} \lambda_i |t^i - t|^2 \log |t^i - t| + \alpha |t-b_0|^2 \log |t-b_0| + q(t)$$

with $q \in P_1$.

Then σ_1 is obtained solving a linear system.

Similar considerations apply for the computation of σ_{k+1} in step k.

(2) From the last remark it appears that each iteration costs the solution of one or two linear systems where the matrices which depend on the knots are different from one step to the other. This could be very expensive to do unless we know something more about the systems to solve.

Here the key remark is that the difference between the original matrix (to compute σ_1) and the matrix used in the computation of σ_k is only given by a rank 1 perturbation. Then we can use this fact to solve the system in step k using the solution in step 0 and the Cholesky factors obtained there. (For details see [35]).

In a recent Ph.D. thesis presented at the University of Wisconsin, M. Villalobos has used a different approach to give an approximate solution to (4.1.1). His approach basically consists in the discretization of the constraints, imposing a regular grid over C and replacing $\mu(t) \geq 0$, $t \in C$ by

$$\mu(g^k) \geq 0 \quad , \qquad g^k \in G \quad ,$$

G: grid over C. (For details, see [36]).

5. Thin Plate Smoothing Splines

In the preceding sections we have considered the solution of some interpolation problems in two dimensions. In all the cases we have assumed that the data are exact and we wanted to obtain a smooth surface passing through the points $(t^i, z_i) \in \mathbb{R}^3$. Our aim in this section is to present the case of noisy data. Thus, in the first part we present the theoretical results giving a characterization of the thin plate smoothing spline and in the second part we give the numerical methods used in the computation of such splines.

5.1. Existence, Uniqueness and Characterization

As in Section 1.2, we assume the model

$$z_i = f(t^i) + \varepsilon_i \ , \qquad i = 1,2,\ldots,n \tag{5.1.1}$$

where $f \in D^{-2}L^2(\mathbb{R}^2)$ is unknown and the ε_i's are supposed to be realizations of independent identically distributed Gaussian variables.

In order to approximate f, we introduce the thin plate smoothing spline of parameter $\lambda > 0$ as the unique solution to (cf. [13], [14])

$$\underset{\mu \in D^{-2}L^2(\mathbb{R}^2)}{\text{minimize}} \left\{ \lambda J(\mu) + \frac{1}{n} \sum_{i=1}^{n} (\mu(t^i) - z_i)^2 \right\} \quad . \tag{5.1.2}$$

As in the one-dimensional case, λ controls the tradeoff between the smoothness of the solution measured by $J(\mu)$ and the approximation to the data measured by

$$\frac{1}{n} \sum_{i=1}^{n} (\mu(t^i) - z_i)^2 \quad .$$

In order to prove the existence and uniqueness of $\sigma_{n,\lambda}$, the solution to (5.1.2), let us first define the $n\times n$ symmetric semi-positive definite matrix Ω as

$$y^T \Omega y = \min_{\substack{\mu\in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i)=y_i,\ i=1,\dots,n}} J(\mu) \quad . \tag{5.1.3}$$

$y^T\Omega y$ represents the value of J at σ_y, the thin plate spline interpolating y_i at t^i, $i = 1,\dots,n$. Given that $y \to \sigma_y$ is linear, the transformation $y \to J(\sigma_y)$ is quadratic hence the existence of Ω.

Now we write (5.1.2) in the following way

$$\min_{y\in\mathbb{R}^n} \min_{\substack{\mu\in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i)=y_i,\ i=1,\dots,n}} \left\{\lambda J(\mu) + \frac{1}{n}\sum_{i=1}^{n} (\mu(t^i)-z_i)^2\right\}$$

$$= \min_{y\in\mathbb{R}^n} \left\{ \min_{\substack{\mu\in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i)=y_i,\ i=1,\dots,n}} \left\{\lambda J(\mu) + \frac{1}{n}\sum_{i=1}^{n} (y_i-z_i)^2\right\}\right\}$$

$$= \min_{y\in\mathbb{R}^n} \left\{\left\{\lambda \min_{\substack{\mu\in D^{-2}L^2(\mathbb{R}^2) \\ \mu(t^i)=y_i,i=1,\dots,n}} J(\mu)\right\} + \frac{1}{n}\sum_{i=1}^{n} (y_i - z_i)^2\right\}$$

$$= \min_{y\in\mathbb{R}^n} \left\{\lambda y^T \Omega y + \frac{1}{n}\sum_{i=1}^{n} (y_i-z_i)^2\right\} \quad . \tag{5.1.4}$$

The solution $\hat{y}$ to this problem is easily found by differentiating and setting the gradient equal to zero. We obtain

$$(n\lambda\Omega + I)\hat{y} = z \quad . \tag{5.1.5}$$

And $\sigma_{n,\lambda}$ is obtained interpolating $\hat{y}_i$ at t^i, $i = 1,\ldots,n$ using a thin plate spline. As in the one-dimensional case, (5.1.5) is only of theoretical use as we shall see later. A more practical result is given by:

<u>Theorem 5.1.1</u>. Let $\{t^1, \ldots, t^n\}$ contain at least one P_1-unisolvent set. Then there exists one and only one solution $\sigma_{n,\lambda}$ to (5.1.2). Moreover, there exist $n+3$ real numbers $\lambda_1, \ldots, \lambda_n$; $\alpha_1, \alpha_2, \alpha_3$ such that $\sigma_{n,\lambda}$ has the expression

$$\sigma_{n,\lambda}(t) = \sum_{i=1}^{n} \lambda_i |t-t^i|^2 \log |t-t^i| + \alpha_1 t_1 + \alpha_2 t_2 + \alpha_3 \quad (5.1.6)$$

where $t = (t_1,t_2)$.

The coefficients $\lambda_1, \ldots, \lambda_n, \alpha_1, \alpha_2, \alpha_3$ are the solutions of the system

$$\begin{aligned} 8\pi n\lambda\lambda_i + \sum_{j=1}^{n} \lambda_j K(t^i-t^j) + \alpha_1 t_1^i + \alpha_2 t_2^i + \alpha_3 &= z_i , \quad i = 1,\ldots,n \\ \sum_{j=1}^{n} \lambda_j t_1^j &= 0 \\ \sum_{j=1}^{n} \lambda_j t_2^j &= 0 \\ \sum_{j=1}^{n} \lambda_j &= 0 \quad . \end{aligned} \qquad (5.1.7)$$

<u>Proof</u>. See [13], [14].

In the next section we use this formulation to compute the smoothing spline. Now we are interested in obtaining an interpretation of $\sigma_{n,\lambda}$ as a filtering process. To do that, let $\rho_1 \le \rho_2 \le \cdots \le \rho_n$ be the eigenvalues of Ω in increasing order and let Q be the orthogonal matrix diagonalizing Ω. Then

$$\Omega = Q \Sigma Q^* \tag{5.1.8}$$

where

$$\Sigma = \mathrm{diag}\,(\rho_1, \ldots, \rho_n) \quad .$$

Then (5.1.5) becomes

$$\tilde{y} = (I + n\lambda\Sigma)^{-1}\tilde{z} \tag{5.1.9}$$

where

$$\tilde{y} = Q^*\hat{y}$$

$$\tilde{z} = Q^*z \quad .$$

Hence

$$\tilde{y}_i = \frac{1}{1+n\lambda\rho_i}\,\tilde{z}_i \quad . \tag{5.1.10}$$

But it has been observed, and proved in some cases [30], [39] that

$$\rho_i \sim C\,\frac{i^4}{n} \quad , \tag{5.1.11}$$

then

$$\tilde{y}_i \sim \frac{1}{1+C\lambda i^4}\,\tilde{z}_i$$

which turns to be again a low-pass Butterworth type filter. For details see [35], [39], [43].

5.2. The Computation of λ and $\sigma_{n,\lambda}$

As in the one-dimensional case, the estimation of λ from the data is a very delicate problem. One of the methods that is becoming the most popular to solve this problem is the Generalized Cross-Validation which amounts to minimizing the GCV function

$$V(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(\sigma_{n,\lambda}(t^i) - z_i)^2}{\left(1 - \frac{1}{n}\mathrm{tr.}(A(\lambda))\right)^2} \tag{5.2.1}$$

where $A(\lambda)$ is given by

$$A(\lambda) = (I + n\lambda\Omega)^{-1} \quad . \tag{5.2.2}$$

Thus, we need an algorithm to compute $\sigma_{n,\lambda}$ and one to compute tr. $(A(\lambda))$. We begin with the computation of $\sigma_{n,\lambda}$.

As we have seen in Th. 5.1.1, the computation of $\sigma_{n,\lambda}$ involves the solution of a linear system in the $n+3$ coefficients $\lambda_1, \ldots, \lambda_n; \alpha_1, \alpha_2, \alpha_3$. The linear system to be solved is given by equations (5.1.7) or, using the notation of Section 2.2, in matrix form

$$\begin{aligned} (8\pi n\lambda + K)\Lambda + E\alpha &= z \\ E^T \Lambda &= 0 \quad . \end{aligned} \tag{5.2.3}$$

Thus, the linear system to be solved has the same form as (2.2.2) and using the same technique, let Q, R be the QR decomposition of E

$$Q^*E = \begin{bmatrix} R \\ 0 \end{bmatrix}$$

where Q^* is $n\times n$ and can be decomposed as

$$Q^* = \begin{bmatrix} Q_1^* \\ Q_2^* \end{bmatrix}$$

and Q_1^* is $3\times n$, Q_2^* is $(n-3)\times n$.

Now setting

$$\tilde{\Lambda} = \begin{bmatrix} \tilde{\Lambda}_1 \\ \tilde{\Lambda}_2 \end{bmatrix} = \begin{bmatrix} Q_1^*\Lambda \\ Q_2^*\Lambda \end{bmatrix} = Q^*\Lambda \quad ,$$

the equations become

$$\begin{aligned} \tilde{\Lambda}_1 &= 0 \\ Q_2^*(8\pi n\lambda I + K)Q_2\tilde{\Lambda}_2 &= Q_2^* z \\ R\alpha &= Q_1^* z - Q_1^* K\, Q_2\tilde{\Lambda}_2 \quad . \end{aligned} \tag{5.2.4}$$

This linear system can be easily solved using Cholesky's decomposition and all the remarks given in the case of interpolation apply to this case.

The computation of tr. $(A(\lambda))$ requires additional work. In [35] we proved that in some cases the eigenvalues of Ω can be well approximated by those of a fourth-order differential operator, but in two dimensions this is too complicated to calculate and a direct evaluation of the eigenvalues is necessary. We first obtain an expression for Ω; to do this we consider the thin plate spline interpolating $y = (y_1,\ldots,y_n)$. Thus its coefficients $\delta_1, \ldots, \delta_n; \beta_1, \beta_2, \beta_3$ satisfy the system

$$\begin{aligned} K\Delta + E\beta &= y \\ E^T\Delta &= 0 \end{aligned}$$

and using the notation

$$\tilde{\Lambda} = Q^*\Delta = \begin{bmatrix} Q_1^*\Delta \\ Q_2^*\Delta \end{bmatrix} = \begin{bmatrix} \tilde{\Lambda}_1 \\ \tilde{\Lambda}_2 \end{bmatrix} \quad ,$$

we get

$$Q_2^* K Q_2 \tilde{\Delta}_2 = Q_2^* y \qquad (5.2.5)$$

$$\tilde{\Delta}_1 = 0 \quad .$$

Now we observe that from Th. 2.1.1 we have

$$\begin{aligned} J(\sigma_y) &= \Delta^T K \Delta \\ &= \tilde{\Delta}^T Q^* K Q \tilde{\Delta} \qquad (5.2.6) \\ &= \tilde{\Delta}_2 Q_2^* K Q_2 \tilde{\Delta}_2 \quad . \end{aligned}$$

Using now (5.2.5) we finally get

$$J(\sigma_y) = y^T Q_2 (Q_2^* K Q_2)^{-1} Q_2^* y$$

and

$$\Omega = Q_2 (Q_2^* K Q_2)^{-1} Q_2^* \quad . \qquad (5.2.7)$$

Finally, we observe that the eigenvalues of Ω are those of $Q_2^* K Q_2$ except for the first three that are equal to zero, then

$$\text{tr.}\ (I + n\lambda\Omega)^{-1} = 3 + \text{tr.}\ (I + n\lambda Q_2^* K Q_2)^{-1} \quad . \qquad (5.2.8)$$

Thus, if we want to compute tr. $(A(\lambda))$ many times, as is necessary when minimizing the GCV function, we first compute the eigenvalues of $Q_2^* K Q_2$ and then use them to compute tr. $(A(\lambda))$. For more details and numerical experiences the reader is referred to [35], [36], [43].

6. Non-variational Techniques

As we have pointed out before, the methods we have exposed here are essentially generalizations of splines in one dimension using the variational properties of those functions. We have chosen this way because of its close relationship with stochastic estimation and because it appears to be at present the only technique used for scattered data (when referring to splines). Many other generalizations are possible, and we will briefly describe two of them in these sections. For a detailed discussion see the references below.

6.1. Tensor Product Splines

The first generalization of one-dimensional splines to many dimensions has been given using tensor products. More precisely, consider a square $[a,b]\times[c,d]$ in $\mathbb{R}^2$ and two partitions, $a = a_1 < a_2 < \cdots < a_n = b$; $c = c_1 < c_2 < \cdots < c_m = d$ of $[a,b]$ and $[c,d]$ respectively. Also let B_i^1, $i = 1, \ldots, n+1$ and B_j^2, $j = 0, \ldots, m+1$; the B-splines based on the knots $\{a_1,a_1,a_1,a_1,a_2,a_3,\ldots,a_n,a_n,a_n,a_n\}$ and $\{c_1,c_1,c_1,c_1,c_2,c_3,\ldots,c_m,c_m,c_m,c_m\}$ respectively.

As we have already said (cf. Section 1.1), $\{B_i^1\}_{i=0}^{n+1}$ is a basis for the space of twice differentiable piecewise cubic polynomials based on the knots $\{a_1, \ldots, a_n\}$.

We now define a space of tensor product splines S on the grid (a_i,c_j), $i = 1,\ldots,n$; $j = 1,\ldots,m$ as the linear space spanned by the tensor product of the basis $\{B_i^1\}_{i=0}^{n+1}$; $\{B_j^1\}_{j=0}^{m+1}$. That is, $s \in S$ if and only if

$$s(x,y) = \sum_{i=1}^{n+1} \sum_{j=0}^{m+1} \alpha_{ij}\, B_i^1(x)\, B_j^2(y) \tag{6.1.1}$$

for $(x,y) \in [a,b]\times[c,d]$ and $\alpha_{ij} \in \mathbb{R}^2$ for $i = 0, \ldots, n+1$; $j = 0, \ldots, m+1$.

Or, in other words,

$$s = \sum_{i=0}^{n+1} \sum_{j=0}^{m+1} \alpha_{ij} B_i^1 \otimes B_j^2 \quad . \tag{6.1.2}$$

It is well known (cf. [1], [9], [26]) that S is a linear space of dimension $(n+2) \times (m+2)$. Thus, the complete determination of s from interpolation conditions requires to give $2(n+m+2)$ additional boundary conditions. A usual choice for these conditions is:

Hermite type

$$\frac{\partial s}{\partial x}(a_1,c_j) = \text{given constant} , \qquad j = 1,\ldots,m$$

$$\frac{\partial s}{\partial x}(a_n,c_j) = \text{given constant} , \qquad j = 1,\ldots,m$$

$$\frac{\partial s}{\partial y}(a_i,c_1) = \text{given constant} , \qquad i = 1,\ldots,n$$

$$\frac{\partial s}{\partial y}(a_i,c_n) = \text{given constant} , \qquad i = 1,\ldots,n$$

$$\frac{\partial^2 s}{\partial x \partial y}(a_1,c_1) = \text{constant} \tag{6.1.3}$$

$$\frac{\partial^2 s}{\partial x \partial y}(a_1,c_m) = \text{constant}$$

$$\frac{\partial^2}{\partial x \partial_{\,}}(a_n,c_1) = \text{constant}$$

$$\frac{\partial^2 s}{\partial x \partial y}(a_n,c_m) = \text{constant} \quad .$$

As in the one-dimensional case these conditions produce optimal convergence rates but as it was already pointed out, they require some additional information on the function that is not always available. (For other choices see [9], [26]).

The functions in the space S are piecewise bicubic. This means that in each subrectangle $[a_i, a_{i+1}] \times [c_j, c_{j+1}]$ the function is a bicubic polynomial, i.e., a function of the form

$$p(x,y) = q_0(x) + yq_1(x) + y^2q_2(x) + y^3q_3(x) \qquad (6.1.4)$$

where q_0, q_1, q_2, q_3 are cubic polynomials.

Moreover, this function $s \in S$ has continuous derivatives $\frac{\partial s}{\partial x}$, $\frac{\partial s}{\partial y}$, $\frac{\partial^2 s}{\partial x \partial y}$, $\frac{\partial^2 s}{\partial x^2}$, $\frac{\partial^2 s}{\partial y^2}$.

Finally, it is interesting to say that the solution of the linear system in α_{ij}, $j = 0,\ldots,m+1$, $i = 0,\ldots,n+1$, can be efficiently performed using very specialized techniques (cf. [1], [9], [26]). For further details the reader is referred to the extensive literature in this area.

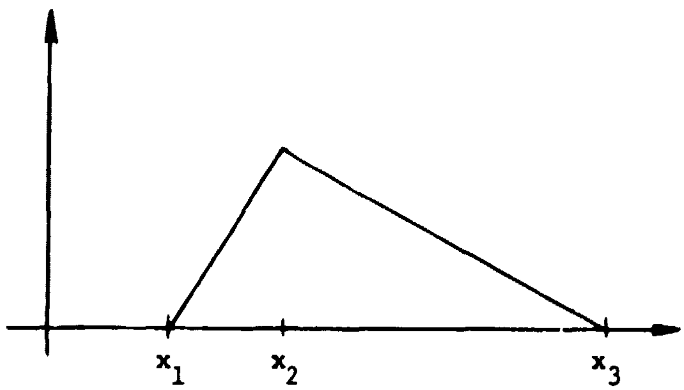
6.2. Multidimensional B-splines

A more recent approach to the problem of multidimensional data has been given by DeVore, Dahmen, Micchelli and others [7], [8], [26]. This new approach is attracting much attention recently, mainly from potential users in finite elements codes.

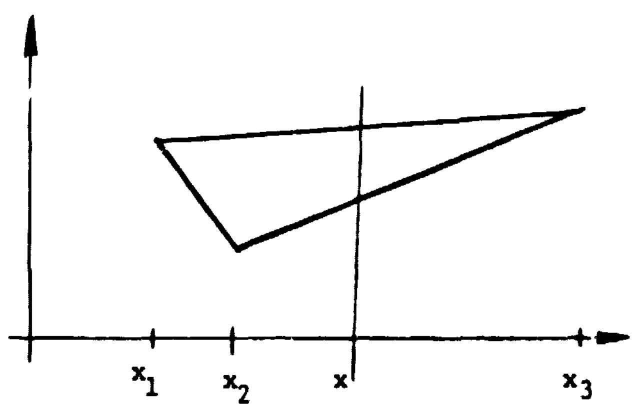
The idea of this method is to generalize the B-splines to several dimensions. The idea is interesting because the two main problems arising with thin plate splines, that is, full system of equations and instability in the evaluation, could be solved when using a local support basis. Unfortunately, until now the theory of multivariate B-splines seems too complicated to be used in practice. For details the reader is referred to [7], [8] and the references therein. Here we only give the basic idea of the construction. To do this we go back to the simplest one-dimensional case, the B-splines of degree 1. These are defined as

$$M(x_1, x_2, x_3; t) = \delta_{x_1,x_2,x_3}(t-x)_+ \qquad (6.2.1)$$

where we have made explicit the dependence of the B-spline on x_1, x_2, x_3, the knots of the spline. A typical plot of M is given below



Of course the use of (6.2.1) to define the multi-dimensional B-spline would require the definition of a multivariate divided difference. Thus, this way seems difficult. However we can still generalize (6.2.1) if we observe that we can give a geometric interpretation of this definition. To do that, let (x_1,y_1), (x_2,y_2), (x_3,y_3) be three points in general position in $\mathbb{R}^2$ (i.e., they form a triangle)



Now for given $x \in \mathbb{R}^2$ define $\tilde{M}(x)$ as the volume in $\mathbb{R}^2$ of the set

$$I_x = \{y \mid (x,y) \in T\} \tag{6.2.2}$$

where

$$T = \text{Triangle with vertices } (x_1,y_1),\ (x_2,y_2),\ (x_3,y_3) \quad .$$

Then

$$\tilde{M}(x) = \mathrm{vol}_{\mathbb{R}^1}(I_x) \quad . \tag{6.2.3}$$

It is clear that $\tilde{M}$ has the same form of $M(x_1,x_2,x_3;\cdot)$ except maybe for its maximum value. Thus (6.2.3) is an alternate definition for M. Clearly this function does not depend on (y_1,y_2,y_3) but only on the volume of T.

The generalization to $\mathbb{R}^S$ is now clear. Let σ be a unit volume simplex in $\mathbb{R}^{S+k}$ and define for $x \in \mathbb{R}^S$ the function

$$M_\sigma(x) = \mathrm{vol}_{\mathbb{R}^k} \{\mu \in \mathbb{R}^k \mid (x,\mu) \in \sigma\} \quad . \tag{6.2.4}$$

This function is a "smooth" piecewise polynomial with support given by the convex hull of the projection of σ in $\mathbb{R}^S$.

Most of the classical results on one-dimensional B-splines can be obtained with these new functions. For example: recurrence relations, integral forms, etc. See [7], [8], [26].

6.3. Conclusion and Comments

Many other subjects might have been included in a complete treatment of two-dimensional or multidimensional splines. The subject is an active area of research and the references given below are only a sample of the extensive literature in the area. The interested reader should consult the symposium on multivariate approximation and related research meetings to have a better idea of the most recent results.

Acknowledgements

This series of lectures at UCLA was made possible by an invitation of Professor A.V. Balakrishnan. The author wishes to thank Professor Balakrishnan for his kind invitation and all the staff of the System Science Department at UCLA for their hospitality during all his stay.

References

[1] Ahlberg, J.H., E.N. Nilson, and J.L. Walsh. The Theory of Splines and Their Applications. New York: Academic Press, 1967.

[2] Aronzajn, N. "Theory of Reproducing Kernels," Trans. Amer. Math Soc. 1950.

[3] Atteia, M. "Fonctions Spline et Noyaux Reprouisants d'Aronszajn-Bergman," R.I.R.O. 1970.

[4] Craven, P. and G. Wahka. "Smoothing Noisy Data with Spline Functions," Numerische Mathematik. 31. pp. 377-403.

[5] Curry, H.B. and I.J. Schoenberg. "On Pólya Frequency Functions. IV: The Fundamental Spline Functions and their Limits," J. Analyse Mathématique. 17, pp. 71-107.

[6] Davis, Philip J. Interpolation and Approximation. New York: Blaisdell, 1963.

[7] Dahmen, W. "On Multivariate B-Splines," SIAM J. on Numerical Analysis. Vol. 17, No. 2, April 1980.

[8] Dahmen, W., R. DeVore and K. Scherer. "Multidimensional Spline Approximation," SIAM J. on Numerical Analysis. 1981.

[9] de Boor, Carl. A Practical Guide to splines. New York: Springer-Verlag, 1978.

[10] del Pino, G., O. Ferreiro and F. Utreras. "Kalman Filter and Spline Functions: A New Bridge Between Numerical and Statistical Analysis." Tech. Rep. U. of Chile. Submitted for Publication.

[11] Demler, C. and C. Reinsch. "Oscillation Matrices with Spline Smoothing," Numerische Mathematik. 24, pp. 375-382, 1975.

[12] Dongarra, J.J., J.R. Bunch, C.B. Moler and G.W. Stewart. Linpack Users Guide. Philadelphia: SIAM, 1979.

[13] Duchon, J. "Interpolation des Fonctions de deux Variables suivant le Principe de la Flexion des plaques Minces," R.A.I.R.O., Analyse Numérique. Vol. 10, No. 12, pp. 5-17, 1975.

[14] Duchon, J. "Splines Minimizing Rotation Invariant Semi-norms in Sobolev Spaces," in: Lecture Notes in Mathematics 571. Heidelberg: Springer-Verlag, 1977.

[15] Duchon, J. "Sur l'Erreur d'Interpolation des Fonctions de Plusieurs Variables par les D^m Splines," R.A.I.R.O. Analyse Numérique. Vol. 12, No. 4, pp. 325-334, 1978.

[16] Dyn, N. and G. Wahba. "On the Estimation of Functions of Several Variables from Aggregated Data," SIAM Journal on Mathematical Analysis. 13, pp. 134-152.

[17] Golub, G., M. Heath and G. Wahka. "Generalized Cross-validation as a Method for Choosing a Good Ridge Parameter," Technometrics. 31, pp. 315-324, 1979.

[18] Hornung, U. "Monotone Spline Interpolation," in: Numerische Methoden der Approximationstheorie, Band 4. Herausgegeben aus L. Collatz, G. Meinardus, H. Werner. Basel: Birkhauser, 1978.

[19] Laurent, P.J. Approximation et Optimisation. Paris: Hermman, 1972.

[20] Laurent, P.J. "An Algorithm for the Computation of Spline Functions with Inequality Constraints." Seminar 335 at IMAG, Grenoble, 1980.

[21] Paihua, L. "Quelques Methodes Numériques pour le Calcul des Fonctions Spline a une et Plusieurs Variables." Thesis, IMAG, Grenoble, 1978.

[22] Paihua, L. and F. Utreras. "Un Ensemble de Programmes pour l'Interpolation de Fonctions par des Fonctions Spline du Type Plaque Mince." R.R. #140, IMAG, Grenoble, 1978.

[23] Prenter, P.M. Splines and Variational Methods. New York: Wiley, 1975.

[24] Reinsch, C. "Smoothing by Spline Functions," Numerische Mathematik. 10, pp. 177-183, 1967.

[25] Schoenberg, I.J. "Contributions to the Problems of Approximation of Equidistant Data by Analytic Functions, Part A," Quarterly of Applied Mathematics. 4, pp. 45-99; Part B, Quarterly of Applied Mathematics. 4, pp. 112-141.

[26] Schumaker, Larry L. Spline Functions: Basic Theory. New York: John Wiley & Sons, 1981.

[27] Schwartz, L. "Sous-Espaces Hilbertiens d'Espaces Vectoriel, Topologiques et Noyaux Assoiés," J. Analyse Mathématiques. 1964.

[28] Schwartz, L. Les Distributions. Paris: Hermman, 1966.

[29] Thomman, J. "Determination et Construction de Fonctions Spline à Deux Variables sur Un Domaine Rectangulaire ou Circulaire." Thesis, Lille, 1970.

[30] Utreras, F.I. "Cross-validation Techniques for Smoothing Spline Functions in One or Two Dimensions," in: Lecture Notes in Mathematics 757. 1979.

[31] Utreras, F.I. "Sur le Choix du Paramètre d'Ajustement dans le Lissage par Fonctions Spline," Numerische Mathematik. 34, pp. 15-28, 1980.

[32] Utreras, F.I. "Convergence Rates for Monotone Cubic Spline Interpolation," Journal of Approximation Theory. Vol. 36, No. 1, 1982.

[33] Utreras, F.I. "Natural Spline Functions: Their Associated Eigenvalue Problem," Numerische Mathematik. 42, pp. 107-117, 1983.

[34] Utreras, F.I. "Smoothing Noisy Data Under Monotonicity Constraints: Existence, Characterization and Convergence Rates." Tech. Rep. U. of Chile, 1983. Submitted for Publication.

[35] Utreras, F.I. "Positive Thin Plate Splines." Technical Report, Texas A&M University, 1984. To appear.

[36] Villalobos, M.A. "Estimation of Posterior Probabilities Using Multivariate Smoothing Splines and Generalized Cross Validation." Thesis, University of Wisconsin at Madison, 1983.

[37] Wahba, G. "Smoothing Noisy Data with Spline Functions," Numerische Mathematik. 24, pp. 383-393, 1975.

[38] Wahba, G. "Practical Approximate Solutions to Linear Operator Equations when the Data are Noisy," SIAM J. on Numerical Analysis. Vol. 14, No. 4, 1977.

[39] Wahba, G. "Convergence Rates of Thin Plate Smoothing Splines," in: Lecture Notes in Mathematics 757. 1979.

[40] Wahba, G. "Surface Fitting with Scattered Noisy Data on Euclidean d-Spaces and on the Sphere," Rocky Mountain Mathematical Journal. 1983.

[41] Wecker, W.E. and C.F. Ansley. "The Signal Extraction Approach to Non-linear Regression and Spline Smoothing." J.A.S.A. 78, pp. 81-89, 1983.

[42] Wegman, E.J. and J.W. Wright. "Splines in Statistics," J.A.S.A. 78, pp. 351-365, 1983.

[43] Wendelberger, J. "Smoothing Noisy Data with Multidimensional Splines and Generalized Cross-validation." Thesis, University of Wisconsin at Madison. 1982.

[44] Wold, S. "Spline Functions in Data Analysis," Technometrics. 16, pp. 1-11, 1974.